Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Hybrid k-Anonymity

Mehmet Ercan Nergiz [a], Muhammed Zahit Gök [b,*]

[a] Computer Engineering Department, Zirve University, Office 446, Gaziantep, Turkey
[b] Computer Engineering Department, Zirve University, Office 107, Gaziantep, Turkey

ARTICLE INFO

Article history:
Received 31 July 2013
Received in revised form
28 February 2014
Accepted 16 March 2014

Keywords:
Privacy
Anonymization
Privacy-preserving databases
k-anonymity
Utility in data publishing

ABSTRACT

Anonymization-based privacy protection ensures that published data cannot be linked back to an individual. The most common approach in this domain is to apply generalizations on the private data in order to maintain a privacy standard such as k-anonymity. While generalization-based techniques preserve truthfulness, relatively small output space of such techniques often results in unacceptable utility loss especially when privacy requirements are strict. In this paper, we introduce the *hybrid generalizations* which are formed by not only generalizations but also the *data relocation* mechanism. Data relocation involves changing certain data cells to further populate small groups of tuples that are indistinguishable with each other. This allows us to create anonymizations of finer granularity confirming to the underlying privacy standards. Data relocation serves as a tradeoff between utility and truthfulness and we provide an input parameter to control this tradeoff. Experiments on real data show that allowing a relatively small number of relocations increases utility with respect to heuristic metrics and query answering accuracy.

1. Introduction

We are experiencing an era in which information is invaluable not only from a research perspective but also from a business perspective. This has led data owners (such as service providers and hospitals) to collect personal information with the hope of turning this data into benefit. In some cases, the potential value of such data is so great, it needs to be outsourced for analysis or it has to be published for research purposes. For example, National Institutes of Health (NIH - official medical research agency of U.S.) expects all sufficiently large projects funded by NIH to include a plan for sharing final research data for research purposes (NIH, 2003).

However, data often contain sensitive information that needs to be kept private such as diagnosis and treatments. Thus sharing it raises every privacy concern. Due to such concerns, data privacy is protected by law in many countries (HIPAA, 2001; European Parliament, 1995). This does not mean, however, data sharing is prohibited. Law protects personal data and data that cannot be linked to an individual identity is not considered personal. Thus, in order to preserve the privacy of individuals, data needs to be properly anonymized (de-identified) before publishing. An anonymization must not only satisfy the underlying privacy requirements but also preserve the utility of the data. Otherwise, it would be difficult to extract useful information from the anonymized data.

NIH explicitly states that just removing uniquely identifying information (e.g., SSN) from the released data is not enough to protect privacy (NIH, 2003):

> In addition to removing direct identifiers, e.g., name, address, telephone numbers, and Social Security Numbers,

* *Corresponding author*. Fax: +90 3422116677.
E-mail addresses: mehmet.nergiz@zirve.edu.tr (M.E. Nergiz), muhammed.gok@std.zirve.edu.tr, mzgok@hotmail.com (M.Z. Gök).
http://dx.doi.org/10.1016/j.cose.2014.03.006
0167-4048/

researchers should consider removing indirect identifiers and other information that could lead to "deductive disclosure" of participants' identities.

Works in Samarati (2001) and Samarati and Sweeney (1998) show that using publicly available sources of partially identifying information (*quasi-identifiers*) such as age, gender and zip-code, data records can be re-identified accurately even if there is no direct identifying information in the dataset. For example, in Table 1, suppose we release T as a private table. Even if T does not contain unique identifiers, an adversary that knows that her 41 years old friend Obi from USA with zip 49001 is in the dataset will be able to identify him as tuple q7.

To prevent identification, many different privacy metrics (Samarati, 2001; Samarati and Sweeney, 1998; Machanavajjhala et al., 2006; Li and Li, 2007; Wong et al., 2006; Nergiz and Clifton, 2009) have been introduced for various adversary models. As an example, *k-anonymity* requires that for each tuple t in the anonymization, there should be at least $k-1$ other tuples indistinguishable with t. Two individuals are said to be indistinguishable if their records agree on the set of quasi-identifier attributes. Various anonymization algorithms have been proposed to achieve the underlying privacy standard. A common feature of these algorithms is that they manipulate the data by using *generalizations* which involves replacing data values with more general values (values that include the meaning of the original value and that may also imply other atomic values, e.g., 'Italy' is changed to 'Europe') so that more tuples will express similar meanings. As an example, suppose the desired privacy standard is 3-anonymity. In Table 1, $T^*_{\mu_1}$ is a 3-anonymous generalization of T. Note that generalizations applied to T create two *equality groups* that contain similar tuples with respect to QI attributes. From the adversary's point of view, tuples with each equality group are indistinguishable from each other. If the data owner releases $T^*_{\mu_1}$ instead of T, Obi can at best be mapped to the white equality group of size 5 and to a set of salaries {18K, 35K, 14K, 25K, 29K}.

A nice feature of generalizations is that unlike perturbation techniques (that apply noise to data cells independently before publishing), generalizations preserve the truthfulness of data. (For example, saying "q7 is from North America" is not wrong. However, saying q7 is of age 40.1 would be wrong.)

However, generalizations result in information loss, thus over-generalization should be avoided as long as the privacy requirements are satisfied. To solve this problem, many heuristics have been designed, however relatively small output space of such techniques often results in unacceptable utility loss especially when privacy requirements are strict (Brickell and Shmatikov, 2008). Preservation of utility still stands as a major problem for generalization-based techniques.

One of the main reasons for over-generalization is the existence of outliers in private datasets. As the neighborhood of the outliers is not heavily populated in the high dimensional domain, it becomes difficult for an anonymization algorithm to generate an equality group of sufficient size. For those algorithms that are vulnerable to outliers, a relatively large group can degrade the overall utility of the whole dataset (Nergiz and Clifton, 2007). For example, in Table T and $T^*_{\mu_1}$, one can consider q4 as an outlier as q4 is older than rest of the equality group and is from West Europe as opposed to East Europe. Similarly tuples q5 and q6 are not very similar to the q7, q8, and q9 and grouping them together creates generalizations of lower quality. Identifying outliers is not an easy task as the definition of an outlier heavily depends on the anonymization algorithm and outliers can be a set of data cells as opposed to a set of tuples.

To address the negative effects of outliers and over-generalization, in this paper, we extend our work in Nergiz et al. (2013) and introduce the *hybrid generalization* technique. Hybrid generalization combines the generalization technique with a *data relocation* mechanism in order to achieve more utilized anonymizations. Data relocation involves changing certain data cells (that act as outliers) to further populate small equality groups of tuples. Over relocation harms truthfulness and localized utility, thus over-relocation should be

Table 1 – T: private table; $\widehat{T}$: a 10%-relocation of T; $T^*_{\mu_1}$, $\widehat{T}^*_{\mu_2}$: 3-anonymous single dimensional generalizations of T and $\widehat{T}$ respectively; $T^*_{\mu_2}$: a single dimensional generalization of T.

T

Id	Age	Nation	Zip	Sal.
q1	12	Greece	47906	13K
q2	19	Turkey	47907	15K
q3	17	Greece	47907	28K
q4	23	Spain	49703	14K
q5	38	Brazil	49705	18K
q6	33	Peru	49812	35K
q7	41	USA	49001	14K
q8	43	Canada	49001	25K
q9	48	Canada	49001	29K

$\widehat{T}$

Id	Age	Nation	Zip	Sal.
q1	12	Greece	47906	13K
q2	19	Turkey	47907	15K
q3	17	Greece	47907	28K
q4	31	Brazil	49703	14K
q5	38	Brazil	49705	18K
q6	33	Peru	49812	35K
q7	41	USA	49001	14K
q8	43	Canada	49001	25K
q9	48	Canada	49001	29K

$T^*_{\mu_1}$

Id	Age	Nation	Zip	Sal.
q1	11-30	EU	4*	13K
q2	11-30	EU	4*	15K
q3	11-30	EU	4*	28K
q4	11-30	EU	4*	14K
q5	31-50	AM	4*	18K
q6	31-50	AM	4*	35K
q7	31-50	AM	4*	14K
q8	31-50	AM	4*	25K
q9	31-50	AM	4*	29K

$T^*_{\mu_2}$

Id	Age	Nation	Zip	Sal.
q1	11-20	E. EU	47*	13K
q2	11-20	E. EU	47*	15K
q3	11-20	E. EU	47*	28K
q4	21-30	W. EU	49*	14K
q5	31-40	S. AM	49*	18K
q6	31-40	S. AM	49*	35K
q7	41-50	N. AM	49*	14K
q8	41-50	N. AM	49*	25K
q9	41-50	N. AM	49*	29K

$\widehat{T}^*_{\mu_2}$

Id	Age	Nation	Zip	Sal.
q1	11-20	E. EU	47*	13K
q2	11-20	E. EU	47*	15K
q3	11-20	E. EU	47*	28K
q4	31-40	S. AM	49*	14K
q5	31-40	S. AM	49*	18K
q6	31-40	S. AM	49*	35K
q7	41-50	N. AM	49*	14K
q8	41-50	N. AM	49*	25K
q9	41-50	N. AM	49*	29K